From Data to Knowledge to Action: Enabling Personalized Education

Beverly Park Woolf
University of Massachusetts-Amherst

Ryan Baker
Worcester Polytechnic Institute

Erwin P. Gianchandani
Computing Research Assoc.

Computing Community Consortium
Version 9: September 2, 2010[1]

As a nation built largely on knowledge – and facing unparalleled twenty-first century challenges – we require our citizens to constantly acquire new skills quickly, to engage in new learning approaches enthusiastically, and to form new learning communities that work well together. Yet most of our classrooms still look like nineteenth and early twentieth century schoolhouses: learning materials (e.g., textbooks and blackboards) have changed little, and teachers use many of the same instructional methods, such as lecturing to passive students or assigning tasks to be solved by individuals. We simply are not cultivating within the next-generation workforce the kinds of understanding and capabilities that we so desperately need in order to be able to tackle the real-world issues of the future. As a result, our students are suffering: national reading tests indicate that almost 90 percent of inner-city fourth graders do not have a basic level of reading proficiency; in international math tests, high school students in Cyprus and South Africa routinely surpass American high school seniors in their results; and nearly half of all minority students drop out of high school.

However, in recent years, education informatics, i.e., an approach to education focused on collecting, mining, and analyzing large data sets about learning, has begun to offer new information and tools to key stakeholders in education, including students, teachers, parents, school administrators, and employers. The resultant data-rich instructional methods offer tremendous promise for transforming education within the U.S. – and can alter how teachers make real-time decisions in classrooms and, consequently, how we educate the next generation of teachers.

Educational data can provide us with an improved understanding of students' knowledge and better assessments of their progress. Data can shed light on key questions in education and psychology, such as the mechanics of learning, how different students (e.g., high- vs. low-achieving, male vs. female, typical vs. those with learning disabilities) respond to different pedagogical strategies, and the overall success and failure of specific teaching approaches. Data can also help us generate models that discern not only what individual students know, but how deeply and richly they know it. Finally, data can enable new ways of finding clusters of children with similar learning styles or difficulties, thereby personalizing the education that we provide.

As an example, the 2010 Knowledge Discovery and Database (KDD) cup featured an educational database as the basis of a worldwide competition to model learning and predict

---

[1] Contact: Erwin Gianchandani, Director, Computing Community Consortium (202-266-2936; erwin@cra.org). For the most recent version of this essay, as well as related essays, visit http://www.cra.org/ccc/initiatives.

outcomes[2]. About one–half million student records within the NSF-supported Pittsburgh Science of Learning Center's "DataShop" – representing the usage of learning software called Cognitive Tutor by hundreds of students (around 50 hours per student) – were made available to global participants. The participants modeled these data and predicted the performance of random students in subsequent problems. Initial results were quite promising. This work is just the beginning of the use of massive databases to make predictive models that ask critically important questions about education: How quickly or slowly do different students learn? What are the underlying factors that make topics easier or harder for students? How should lesson design and curriculum be modified? How can we respond when students become disengaged?

In this paper, we describe how data analytics approaches have the potential to dramatically advance instruction for every student and to enhance the way we educate our children. The Internet, intelligent environments, and rich interfaces (including sensors) allow us to capture much more data about learners than ever before – and the quantities of data are growing at a rapidly accelerating rate. Coupled with recent advances in data mining, machine learning, and reasoning, as well as rapid rises in computing power and storage, we are transforming our ability to understand increasingly large, heterogeneous, noisy or incomplete datasets collected during learning. Below we illustrate the *data* → *knowledge* → *action* paradigm in the context of education: instructional systems repeatedly observe how students react and store these data into public repositories; these data are then analyzed to infer new knowledge about learning; and this information drives real-time decision-making.

Data analytics as a driver

Data analytics approaches allow us to optimize teaching tools, personalize education, improve the measurement and value of student and teacher assessments, etc. Here we describe specific research directions underlying these advances.

Managing large educational databases. Greater investment is needed to build open repositories for storing and sharing educational data and to develop algorithms particularly adapted to the educational domain and its unique characteristics. How do we effectively store, manage, make available and analyze data for different purposes and stakeholders? Educational data mining enables educators to "look at" diverse repositories of data wherever they may be and with sufficient processing power for any desired algorithm to process the data. Vast amounts of data on large numbers of students are stored in public repositories and are made available (in properly privatized and analyzed form) to the broader research community. These data are currently available for some types of learning systems (such as intelligent tutoring systems), but they need to be made available for the full diversity of contexts and systems in which students learn. Steps should be taken to make data management of enormous files possible, perhaps by merging the capabilities of file systems to store and transmit data from experiments, using logical organization of files, and employing specific query languages that enable analytic operations. Metadata should be made available describing each experiment and the data it produced. The full power of relational databases will be available to allow effective interactions with the data. Interfaces will be available along with toolkits for purposes of visualizing and plotting the data. Methods are needed to quickly label educational data in support of supervised learning

[2] https://pslcdatashop.web.cmu.edu/KDDCup/.

algorithms that are used to develop prediction models that can detect differences in student engagement, motivation, meta-cognition, and learning strategy. Research is needed to ensure that metadata descriptions mean the same thing when used with different systems. One approach is to define central ontologies of learning objectives used to organize and index the systems. An alternative is the folksonomy approach, where structure emerges from decentralized tagging. Both approaches are in use and their relative merits are being evaluated. Bringing the results of these analyses into educational systems research and recommending an educational systems architecture are important challenges in the current time frame.

Predictive models. Predictive models leverage years of collecting heterogeneous instructional data, generating context-sensitive instruction, real-time feedback, and adaptive problems, and thus serving as hugely valuable accessories to teachers. Models will be able to *predict* students' learning styles, enabling presentations and content to be tailored to individual students in an engaging and efficient manner. Researchers have already constructed models from data collected from thousands of K-12 students using computer-based instructional tools in their schools. These models have been used to predict which educational material will be most effective for each student. Predictive models will be capable of forecasting student engagement, and recommending actions to enhance engagement or to re-engage students. Moreover, predictive models will advise designs for distributing and combining scarce expert pedagogy with always-available online content and tutoring. Finally, they will enable lagging students to catch up, if necessary, in private and highly supportive ways, working at their optimal pace, bringing learners lost to education back into learning.

Adaptive systems. Students have a variety of learning needs (e.g., exceptional students learn beyond their age group, special needs students require accommodations, etc.). Yet educational software is often built for the average student, not for advanced students or slow learners. These inflexible systems are often let loose in constantly changing environments (e.g., through the Web) under conditions that cannot be predicted. This approach is limited and shortsighted because the expertise of software authors often has gaps. Authors have incomplete knowledge of domains and constrained knowledge of students and pedagogy. Consequently, portions of their software remain fossilized, incomplete and require human intervention. Developing analytical techniques offers an opportunity to enable a system to adapt to new student populations. Machine learning techniques can enable a system to acquire knowledge about distinct student groups and add that to its original capabilities. Based on experience with prior populations, adaptable software can reason "outside" the original variables provided by the author.

User models. Computer modeling techniques automatically augment user models or representations of what learners know, their ability to learn (meta-cognition) and their affect (e.g., frustrated, bored, engaged). When and how was knowledge learned? What pedagogy worked best for a given learner? Observations of students' past behavior provide training examples that form models designed to predict future actions. These techniques have been used to group students into communities or stereotypes (e.g., high/low achievers in mathematics).

Increased generality. Both practical and theoretical issues are addressed by data analytics techniques for educational systems. Theoretical issues include increased generality, learning about human learning, and reasoning about uncertainty. For example, intelligent instructional

systems lack the generality that science requires of its theories and explanations. Because these systems might be ported to new environments and function under new requirements, general principles about their knowledge and reasoning can help expand them and transfer their functionality to new domains. Consider a college teacher who finds an intelligent system for teaching high school algebra on the Web. This teacher wants to use the tutor for teaching algebra to adults. How will the system identify examples and hints that work best with college students? Can the system extend its teaching domain, perhaps to teach pre- calculus based on its ability to teach algebra? General principles such as these might allow the system to be expanded across multiple students and disciplines.

Other **important approaches to improve education through the use of technology** include:

- Applying variants on Bayesian knowledge tracing that are increasingly used to study a wide variety of constructs;
- Developing better tools for supporting statistical analysis of the differences between data-mined models and methods to generalize data-mined models across contexts;
- Bringing together data miners and psychometricians, to assess the possible benefits that occur from the integration of machine learning and psychometric methods;
- Integrating the results of one model into a second model (e.g., models of learning have been key components in models of other constructs such as how people game systems);
- Researching "discovery within models," in which a machine-learned model of a construct is developed and then utilized in a broader data set, in conjunction with other models or other measures (e.g., survey measures), in order to understand the associations between the constructs studied; and
- Determining how models and model-creation software can be made available for broader use.

Ultimately, by synthesizing, analyzing, and distributing data and knowledge about education to a variety of stakeholders, we improve the odds that learners will succeed. Consider, for example, assessment information. Young learners can benefit from their parents being informed about learning deficiencies and providing additional help or motivation. Similarly, teachers can benefit from seeing a summary of areas of student strengths and weaknesses, which in turn prompt them to immediately alter their teaching methods. Data analytics approaches are thus vital to ensuring the success of our education system in the twenty-first century.

Important considerations

The variety and volume of data collected and the potential use of these data to improve education will continue to grow. While the potential benefits are great, several challenges remain.

Student privacy. Data security and privacy must be addressed to achieve the greatest benefits while protecting students' civil liberties. These issues are central and will require a modernization of existing policies for collecting and using data about individuals. Importantly, there is a key role for technology as well as political process in managing the tradeoff between privacy and the benefits of collecting and using data.

Social constructs in education. Clearly education reform requires a "socio-technical" solution, recognizing the very large impact of social constructs in education. Technological development in school settings is not immediate, and we need to be cognizant of the social, political and economic constraints on school administrators and institutions.

Social learning. We also need to support social learning communities so that they can flourish without requiring that participants or educators have technology skills. Data mining can help identify Internet objects available for shared social learning and objects of conversation (e.g., a shared graph).

The need for Federal investment

Developing and deploying the kinds of tools and technologies described above – and understanding their eventual value and impact – requires substantial Federal investment. The National Science Foundation (NSF) is leading the way, having recently established a multi-disciplinary Cyberlearning program[3] to harness the transformative potential of advanced learning technologies across the education enterprise. NSF's Directorates for Computer and Information Science and Engineering (CISE), Education and Human Resources (EHR), and Social, Behavioral, and Economic Sciences (SBE), as well as its Office of Cyberinfrastructure (OCI) are all part of this highly collaborative initiative. The Cyberlearning program is much broader than the education informatics work described above; it funds research into experiences for engaging users with content, with people who can be mentors, with imaginary worlds, with invisible phenomena, etc. – the kinds of things that will draw learners in and help them see purpose in their learning and sustain their engagement over time. But data analytics is a critical component of the Cyberlearning program, and success in this area hinges upon the program being continued as a large-scale, multi-disciplinary, Foundation-wide effort for many years to come.

NSF's Cyberlearning program also serves as a compelling model for the kinds of initiatives that other mission-critical agencies should develop and pursue. For example, the President's FY 2011 budget request for the National Institutes of Health (NIH) includes about $824 million in training. NIH currently supports clinical training and postdoctoral research fellowships, and it is likely to invest in improved education of patients and care providers in the years ahead, as the challenges of chronic diseases like cancer and diabetes mount. Similarly, the Department of Defense (DoD) spends billions of dollars on training every year, including maintaining mission and tactical readiness, enhancing special operations, and strengthening intelligence and security capabilities. **NSF and DoD should establish programs analogous to NSF's Cyberlearning programs that bring together computer scientists, social scientists, and leading domain experts to identify new technology-based innovations and strategies for instructing the next generation – of clinicians, researchers, patients, war fighters, etc. Of critical importance in these initiatives will be the work in education informatics. For example, the NIH program must make sense of increasing amounts of mobile data to determine which approaches are more or less effective in teaching patients wellness strategies. Ideally, the sizes, lengths, and overall costs of these programs should scale according to the needs of the respective agencies.**

[3] http://www.nsf.gov/about/budget/fy2011/pdf/06-CISE_fy2011.pdf.

Finally, it is imperative for these Federal agencies to collaborate with the Department of Education, particularly the Institute for Educational Studies (IES), much as NSF has already begun doing. IES is responsible for generating rigorous and relevant evidence on which to ground education practice and policy, and the results of the basic research initiatives described above must be communicated broadly to ensure that the appropriate systems are deployed in the appropriate settings under the appropriate time constraints and guidelines. **IES itself provides funding for research – at $200 million – and its budget, particularly collaborative initiatives with NSF and the other mission-critical agencies described above, must continue to grow with the Federal investment in education data analytics, and, more broadly, education technology and cyberlearning.**

The road ahead

Areas like data mining, machine learning, and predictive modeling are increasingly well researched, but they have not yet been combined in large scale, or optimized specifically for education. Nevertheless, these approaches have the potential to radically alter how we harness the deluge of scientific and learning data, monitor our instructional capabilities, and raise new issues, such as dynamic student assessment, personalized feedback, and lifelong learning. New opportunities exist to analyze vast new educational data sets that contain elements of learning, affect, motivation, social interaction, and longitudinal – indeed lifelong – patterns of learning and engagement that will lead to transforming American education. Continued progress in these and other areas of education technology – in the form of Federal investment supporting interdisciplinary teams of computing, social science, and education researchers, from K-12 to college to domain-specific learning – is essential in order to improve the quality and success of education, and ensure America's competitiveness in the twenty-first century global economy.

For citation use: Woolf B. P., Baker R., & Gianchandani E. P. (2010). *From Data to Knowledge to Action: Enabling Personalized Education*: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association.
http://cra.org/ccc/resources/ccc-led-whitepapers/